Weather inputs to hydrological / hydrogeological models

Richard Chandler and Chiara Ambrosino
r.chandler.ucl.ac.uk, c.ambrosino@ucl.ac.uk

Department of Statistical Science, University College London

HYDEF progress meeting, Wallingford, 27th June 2012

Progress on project

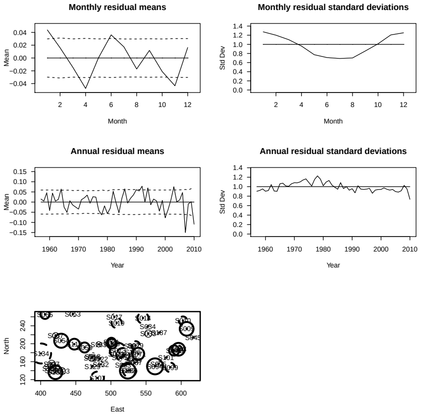
- Software development ongoing
- Daily weather generator: fitting of joint mean-variance model now available (needed for realistic simulation of many weather variables e.g. pressure, temperature):

$$
\begin{aligned}
Y_{st} &\sim N\left(\mu_{st}, \sigma^2_{st}\right) \\
\mu_{st} &= \beta_0 + \sum_{i=1}^{p} \beta_i x_{st}^{(i)} \\
\log \sigma^2_{st} &= \gamma_0 + \sum_{i=1}^{q} \gamma_i z_{st}^{(i)}
\end{aligned}
$$

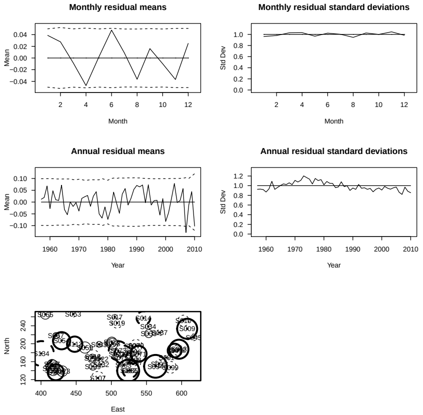
- Pressure model developed for Thames
- Next three months:
  - Software for fitting multivariate models complete
  - Preliminary multivariate model development done for Thames

Example: pressure modelling for Thames

Model with constant variance



Joint mean-variance model



Weather inputs to models: preliminaries

- In HYDEF, (sub-)daily weather data are inputs to hydrological / hydrogeological models
- Basic setup: (deterministic) model produces outputs $\mathbf{y}^*$ as function of inputs $\mathbf{x}^*$ and parameters θ:

$$\mathbf{y}^* = f(\mathbf{x}^*, \theta) .$$

- Models & measurements are imperfect: need to acknowledge discrepancy between model output $\mathbf{y}^*$ and observation $\mathbf{y}$:

$$\mathbf{y} = \mathbf{y}^* + \varepsilon = f(\mathbf{x}^*, \theta) + \varepsilon \ .$$

Models: requirements and uses

- Parameter vector θ often unknown & must be estimated — calibration
- Given θ and inputs $\mathbf{x}^*$, determine outputs $\mathbf{y}^*$ or observations $\mathbf{y}$ — simulation

Question

What if available weather inputs $\mathbf{x}$ are not the same as the required $\mathbf{x}^*$? Possible reasons:

- $\mathbf{x}$ is usually either station data or derived products (e.g. reanalysis)
- $\mathbf{x}^*$ often gridded values / complete records

More reasons why $\mathbf{x} \neq \mathbf{x}^*$

- Problems with station data:
  - Short records (particularly when simultaneous records needed)
  - Spatially inhomogeneous sampling
  - Inhomogeneities / inconsistencies due to observer practice, instrumentation, changing environment, station moves, ...
  - Errors / artefacts due to equipment failure, human / animal interference, transcription error, postprocessing, ...
  - Not all required variables recorded routinely (e.g. for evapotranspiration calculations)
    - Challenge to modellers: please be realistic in your input requirements!

More reasons why $\mathbf{x} \neq \mathbf{x}^*$

- Problems with station data:
  - Short records (particularly when simultaneous records needed)
  - Spatially inhomogeneous sampling
  - Inhomogeneities / inconsistencies due to observer practice, instrumentation, changing environment, station moves, ...
  - Errors / artefacts due to equipment failure, human / animal interference, transcription error, postprocessing, ...
  - Not all required variables recorded routinely (e.g. for evapotranspiration calculations)
    - Challenge to modellers: please be realistic in your input requirements!
- Problems with derived products:
  - Many derived from station data $\Rightarrow$ inherit problems above
  - Most rely on models / algorithms — additional uncertainties / imperfections introduced here

$\mathbf{x} \neq \mathbf{x}^*$: implications for simulation

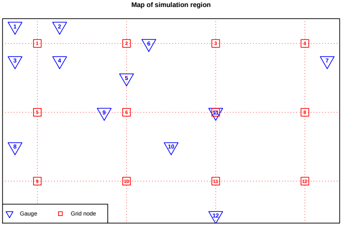
- Common practice: take 'best estimate' as proxy for $\mathbf{x}^*$ e.g. gridded data products
- Many popular products based on some form of interpolation:
  - Inverse distance weighting
  - Kriging
  - etc.

$\mathbf{x} \neq \mathbf{x}^*$: implications for simulation

- Common practice: take 'best estimate' as proxy for $\mathbf{x}^*$ e.g. gridded data products
- Many popular products based on some form of interpolation:
  - Inverse distance weighting
  - Kriging
  - etc.
- But:
  - Interpolated values are smoothed $\Rightarrow$ variability reduced (affects, e.g., extremes)
  - Interpolation introduces artificial inhomogeneities e.g. due to different distances from nearest neighbouring gauges
  - Interpolation gives false impression of reduced uncertainty . . .
- Similar criticisms apply to other forms of 'best estimate'

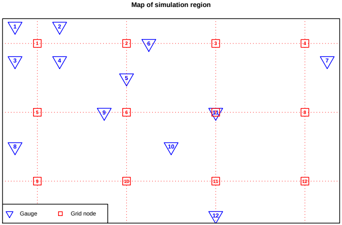
Example: simulation experiment

- Simulate 30-year sequences at 12 locations (blue triangles):
  - Multi-site generalized linear model (GLM) used: identical structure at all sites
  - Sequences 'typical' of SE England
  - Spatial scale: $\sim 75\%$ of days have sites all wet or all dry, wet-day inter-site correlations ~ 0.6–0.8.



Example: simulation experiment

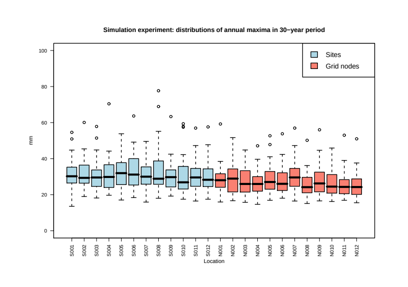
- Simulate 30-year sequences at 12 locations (blue triangles):
  - Multi-site generalized linear model (GLM) used: identical structure at all sites
  - Sequences 'typical' of SE England
  - Spatial scale: $\sim 75\%$ of days have sites all wet or all dry, wet-day inter-site correlations ~ 0.6–0.8.



- Use kriging to create gridded daily dataset from simulations
- Regular grid: 12 nodes (red squares)
- Compare annual maxima / GEV return levels for original & gridded data

Results of simulation experiment

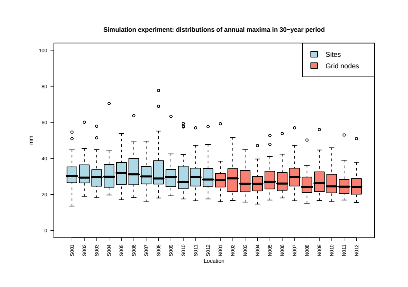
Distributions of annual maxima, and pooled return level estimates



Return period	Estimate (mm) Original	Gridded
10 yr	44.0	38.0
50 yr	57.8	49.4
100 yr	63.9	54.4

Results of simulation experiment

Distributions of annual maxima, and pooled return level estimates

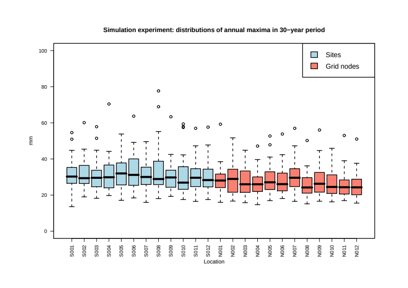


Return period	Estimate (mm) Original	Gridded
10 yr	44.0	38.0
50 yr	57.8	49.4
100 yr	63.9	54.4

- Maxima for gridded data are smaller and less variable
- Gridding reduces return level estimates by $\sim 15\%$

Results of simulation experiment

Distributions of annual maxima, and pooled return level estimates



Return period	Estimate (mm) Original	Estimate (mm) Gridded
10 yr	44.0	38.0
50 yr	57.8	49.4
100 yr	63.9	54.4

Actual return periods for gridded estimates: 5, 19 and 34 years

- Maxima for gridded data are smaller and less variable
- Gridding reduces return level estimates by $\sim 15\%$

An alternative: multiple imputation

- Imputation $=$ sampling missing data from conditional distribution given available observations
- Multiple samples quantify uncertainty due to missing data
- Interesting ideas emerging for visualisation of multiple imputations

An alternative: multiple imputation

- Imputation = sampling missing data from conditional distribution given available observations
- Multiple samples quantify uncertainty due to missing data
- Interesting ideas emerging for visualisation of multiple imputations

Provocative proposal (with support from statistical community)

- Data product creators should routinely provide multiple samples . . .

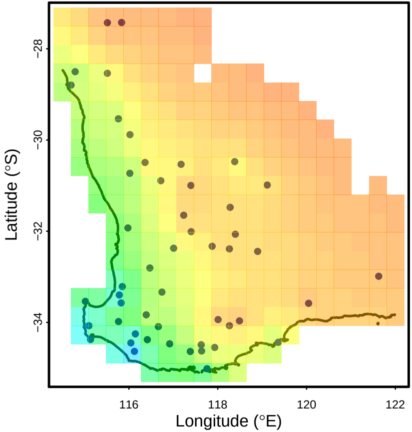
An alternative: multiple imputation

- Imputation = sampling missing data from conditional distribution given available observations
- Multiple samples quantify uncertainty due to missing data
- Interesting ideas emerging for visualisation of multiple imputations

Provocative proposal (with support from statistical community)

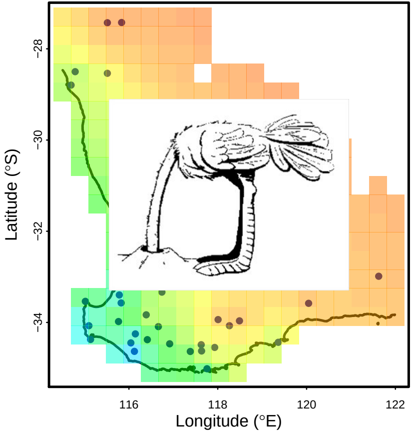
- Data product creators should routinely provide multiple samples ...
- and should **NOT** provide a 'best' value

Example: gridded precipitation



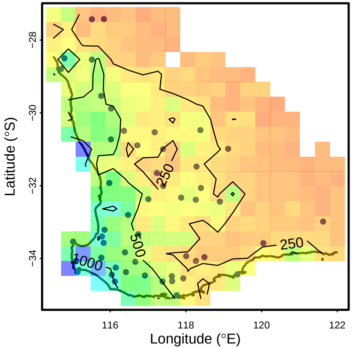
- Area: south-west Western Australia
- Observations: rainfall totals for May–October 2009, from 51 stations
- Requirement: rainfall totals on a fine regular grid
- Interpolation (kriging) yields ‘best’ estimate
- But estimated field doesn’t look like precipitation!

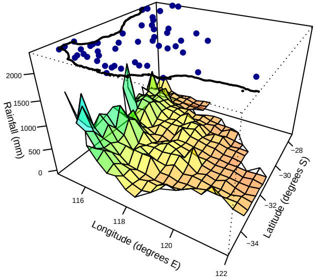
Example: gridded precipitation



- Area: south-west Western Australia
- Observations: rainfall totals for May–October 2009, from 51 stations
- Requirement: rainfall totals on a fine regular grid
- Interpolation (kriging) yields 'best' estimate
- But estimated field doesn't look like precipitation!
- Use at your own risk ...

Visualising multiple samples: the user interface?





(idea & sampling algorithm due to Adrian Bowman, University of Glasgow)

$\mathbf{x} \neq \mathbf{x}^*$: implications for calibration

- Recall: goal of calibration is to identify appropriate value of parameters θ
- Given observations $\mathbf{y}$ and (perfect) inputs $\mathbf{x}^*$, calibration usually performed by optimising some objective function:

$$\hat{\theta} = \arg\min_\theta Q(\theta; \mathbf{y}, \mathbf{x}^*), \text{ say.}$$

- $Q(\cdot)$ chosen to penalise differences between observations $\mathbf{y}$ and model outputs $\mathbf{y}^* = f(\mathbf{x}^*, \theta)$:
  - Weighted or unweighted least-squares criterion
  - Negative log-likelihood
  - Etc.

Example: least-squares fitting of a straight line

- Model: $\mathbf{y}^* = (y_1^* \ \ldots \ y_n^*)'$ etc.,

$$y_i^* = \alpha + \beta x_i^* \ .$$

- Parameter vector is $\theta = (\alpha \ \beta)'$.
- Least-squares objective function is

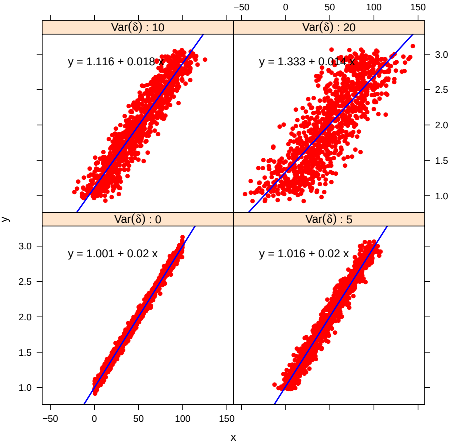
$$Q(\theta; \mathbf{y}, \mathbf{x}^*) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i^*)^2 \ ,$$

which gives accurate and unbiased estimates of θ under general conditions if n is large.

What if $\mathbf{x} \neq \mathbf{x}^*$ and we minimise $Q(\theta; \mathbf{y}, \mathbf{x})$ instead of $Q(\theta; \mathbf{y}, \mathbf{x}^*)$?

$\mathbf{x} \neq \mathbf{x}^*$: effect on fitting a straight line

- Model: $y_i^* = 1 + 0.02x_i^*$ but only have $x_i = x_i^* + \delta_i$ for $x_i^* = 1, \ldots, 100$
- Take $\delta \sim N(0, \sigma^2)$ with $\sigma^2 = 0, 5, 10, 20$
- Take $y_i \sim N(y_i^*, 0.05^2)$
- **NB** intercept increases & slope decreases as $\mathbf{x}$ diverges from $\mathbf{x}^*$
- Result holds generally for linear regression models ('regression dilution bias'); similar issues for more complex models



Confronting the calibration problem

- Previous examples show that ignoring errors / uncertainty in inputs can lead to biased / non-physical model calibration
- How to address this? Ideas from statistical literature:
  - SIMEX (SIMulation-based EXtrapolation) — add extra noise to inputs and then extrapolate back to zero noise
  - Bayesian methods — represent all quantities explicitly (computationally challenging)
  - Estimating equations — cheap and cheerful?

Estimating equations in 2 minutes

- Idea: calibration often done by solving estimating equation $\mathbf{g}(\theta; \mathbf{x}, \mathbf{y}) = \mathbf{0}$ (**NB** objective function optimisation covered by this — $\mathbf{g}(\cdot)$ is gradient vector).
- If 'target' value of θ is θ_0 i.e. $\mathbf{y}^* = f(\mathbf{x}^*, \theta_0$ then unbiased estimating equation has $\mathbb{E}[\mathbf{g}(\theta_0; \mathbf{x}, \mathbf{y})] = \mathbf{0}$.
  - Expectation implies probability distribution — uncontroversial if multiple sets of observations $(\mathbf{x}, \mathbf{y})$ are possible given same set of model quantities $(\mathbf{x}^*, \mathbf{y}^*)$.
- Under fairly general conditions, unbiased estimating equations lead to decent estimators of θ in large samples.
- Details: Jesus & REC, *Interface Focus* 2011.
- Implication: bias-correct the estimating equation for $\mathbf{x} \neq \mathbf{x}^*$, and correction of $\hat{\theta}$ will follow.

Example: bias-correcting the linear regression model (I)

- Given $y_i^* = \alpha_0 + \beta_0 x_i^*$, $y_i = y_i^* + \varepsilon_i$, $x_i = x_i^* + \delta_i$, $\delta_i \sim N(0, \sigma^2)$, might consider minimising least-squares objective function as before
- Leads to estimating equation

$$\mathbf{g}(\theta; \mathbf{x}, \mathbf{y}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\theta) = \mathbf{0} \Rightarrow \hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \; ,$$

$$\text{where } \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \text{and} \quad \mathbf{X}' = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$$

- At 'target' value $\theta_0 = (\alpha_0 \; \beta_0)'$, can show that

$$\mathbb{E}[\mathbf{g}(\theta_0; \mathbf{x}, \mathbf{y})] = -\begin{pmatrix} 0 & 0 \\ 0 & n\sigma^2 \end{pmatrix} \theta_0 = -\mathbf{V}\theta_0 \text{ (say) } \neq 0 \; ,$$

so estimating equation is biased.

Example: bias-correcting the linear regression model (II)

- Previous result shows that

$$\mathbb{E}\left[\mathbf{g}\left(\theta_0;\mathbf{x},\mathbf{y}\right)\right] = \mathbb{E}\left[\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\theta_0\right)\right] = -\mathbf{V}\theta_0 \ .$$

- Suggests modifying the estimating equation to

$$\tilde{\mathbf{g}}\left(\theta;\mathbf{x},\mathbf{y}\right) = \mathbf{X}'\left(\mathbf{y}-\mathbf{X}\theta\right) + \mathbf{V}\theta = \mathbf{0} \ ,$$

  which is unbiased.

- Corresponding estimator is $\tilde{\theta} = \left(\mathbf{X}'\mathbf{X}-\mathbf{V}\right)^{-1}\mathbf{X}'\mathbf{y}$ instead of least-squares estimator $\hat{\theta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$
- **NB** need to know **V** — but this should come with uncertainty assessment for weather inputs
- Various refinements possible; standard errors available etc.

Example: bias-correcting the linear regression model (III)

Simulation example revisited

Setup as before, but now with

$n = 1000, x_1^* = 0.1, x_2^* = 0.2, \ldots, x_{1000}^* = 100$ to show effect more clearly.

Target values: $\alpha_0 = 1$, $\beta_0 = 0.02$.

$\text{Var}(\delta)$	$\hat{\alpha}$	$\hat{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
0	1.001	0.020	1.001	0.020
5	1.016	0.020	0.986	0.020
10	1.116	0.018	1.015	0.020
20	1.333	0.014	1.022	0.020

Discussion points

- How do you feel about working with multiple samples of weather inputs?
- Simulation under uncertain weather inputs is definitely an issue for HYDEF. What about calibration?
- What is done in the hydrological / hydrogeological communities about this at present?
- Any better ideas?